

How Web 1.0 Fails: The Mismatch Between Hyperlinks and Clickstreams

Lingfei Wu*

Robert Ackland†

January 31, 2012

Abstract

The core of the Web is a hyperlink navigation system collaboratively set up by webmasters to help users find desired websites. But does this system really work as expected? We show that the answer seems to be negative: there is a substantial mismatch between hyperlinks and the pathways that users actually take. A closer look at empirical surfing activities reveals the reason of the mismatch: webmasters try to build a global virtual world without geographical or cultural boundaries, but users in fact prefer to navigate within more fragmented, language-based groups of websites. We call this type of behavior “preferential navigation” and find that it is driven by “local” search engines.

*Department of Media and Communication, City University of Hong Kong.

†Corresponding Author. Australian Demographic and Social Research Institute, The Australian National University. Postal address: Coombs Building (no. 9), The Australian National University, ACT 0200, AUSTRALIA. TEL: +61 02 6125 0312; FAX: +61 02 6125 2992, robert.ackland@anu.edu.au

1 Introduction

Invented by Tim Berners Lee in 1991, the World Wide Web is regarded as the “largest human information construct in history” (<http://webscience.org/webscience.html>). The Web is commonly understood to have had three overlapping phases of development or eras. Under Web 1.0, webmasters provide content that is consumed by users, while Web 2.0 blurs the distinction between webmasters and users, with blogging tools, social network sites (e.g. Facebook) and micro-blog services (e.g. Twitter) enabling non-technical people to both produce and consume content (“prosumption”)[19, 18, 16]. Web 3.0, or the Semantic Web, involves technologies that make the Web more machine-readable, leading to a “web of data”, which is an evolution of the Web 1.0 “web of documents” [17].

A common feature of all three phases is the use of technologies to help people find web content. With Web 1.0, and to a lesser extent Web 2.0, the core enabling technology is the hyperlink, which allows users to efficiently move around the Web, while Web 3.0 envisages automated agents finding content on behalf of users by drawing on users’ browsing habits.

The importance of hyperlinks to the Web has led to a large amount of research, with applied physics research into properties of hyperlink networks (e.g. power laws) and models that might explain their emergence [4], social scientific studies on the sociological motivations behind hyperlink creation [1] and the political implications of power laws on the Web [8], and computer science research into how hyperlink structures can be used to improve web search [10, 13]. These studies assume that user navigation is completely based on hyperlinks [13, 8], however, recent studies show that users also use search engines and bookmarks to facilitate web surfing [14, 12, 11].

The present paper uses a novel dataset to investigate the extent to which web navigation is based on hyperlinks. We construct clickstream and hyperlink networks comprised of the same 980 websites. In the clickstream network, a directed weighted edge between websites i and j indicates the percentage of global Web users who visited website i and then immediately visited website j . The clickstream network thus shows the pathways that people are taking as they navigate the Web. In the hyperlink network, a directed unweighted edge between websites i and j indicates that i hyperlinks to j , and hence the hyperlink network shows the pathways

that webmasters are creating for users.

Our analysis reveals a substantial mismatch between the hyperlink and clickstream networks, allowing us to conclude that in navigating the Web, users tend to create their own pathways rather than following hyperlinks laid out by webmasters. This mismatch between hyperlinks and clickstreams reveals different preferences of web masters and users: while webmasters work collaboratively to build a fully-connected online society, users in fact only navigate within the fragmental parts of the Web that they favor, a behavior which we term “preferential navigation”.

2 Data and methods

We selected the top 1,000 websites according to Google’s traffic statistics in November 2010 (<http://www.google.com/adplanner/static/top1000/>). We then used Alexa (<http://www.alexa.com>) to retrieve the daily *traffic* to these websites (which is averaged over three months) and also the daily *clickstreams* between them. The sum of clickstreams to a given site will be less than or equal to the traffic to that site, since clickstreams only refer to visits from the set of 1,000 websites, while traffic is *all* visits to the site. According to Google’s statistics, these 1,000 websites account for more than 97% of global Web traffic during the period of the data collection. The clickstream network contains 12,008 directed and weighted edges, where an edge between websites i and j indicates the percentage of global Web users who visited j immediately after visiting i .

We then used VOSON (<http://voson.anu.edu.au/>), which is software for hyperlink network construction and analysis created by one of the authors, to construct a hyperlink network where a directed and unweighted edge between websites i and j indicates that i contains a hyperlink to j . The hyperlink network contains 15,907 edges.

Twenty sites were dropped due to a lack of data and thus our analysis is for the remaining 980 sites. Further, as Alexa reports a maximum of ten largest inbound and outbound clickstreams for each website, and the version of the VOSON web crawler that was used only collected a maximum of 1,000 outbound hyperlinks for each site, the two constructed networks necessarily

do not include all of the clickstreams and hyperlinks between these 980 sites. It is important to note that our data do not allow us to know exactly how a person navigates from website i to j : navigation may occur either through a hyperlink, a search engine, or the user typing the URL into the browser (or equivalently, following a bookmark).

3 Results

The hyperlink network and the clickstream network are shown in Fig.1, with edges that are common to both networks drawn in red. We now explore differences between the two networks and the origins of the differences.

3.1 The mismatch between hyperlinks and clickstreams

The fact that the clickstream and hyperlink networks are comprised of the same nodes allows us to check for overlaps between the two sets of directed edges. It is observed that only 2,580 out of the 15,907 hyperlinks overlap with the 12,008 clickstreams, meaning that a large proportion of hyperlinks are “useless” in the sense that they connect sites that did not exchange any traffic during the data collection period. The clickstreams transported by hyperlinks only accounts for 33% of the sum of all clickstreams. This percentage gives an upper bound of the hyperlink-moderated clickstreams, since we collect more hyperlinks than clickstreams for each website. In other words, the actual proportion clickstreams driven by hyperlinks is likely to be even smaller.

3.2 The story behind the mismatch: Different preferences of webmasters and users

To investigate the origins of the observed mismatch between the hyperlink and clickstream networks, we simulated user navigation in both networks. The simulation is inspired by the sandpile model [2, 3] and the label-propagation algorithm [15]. In the simulation, a group

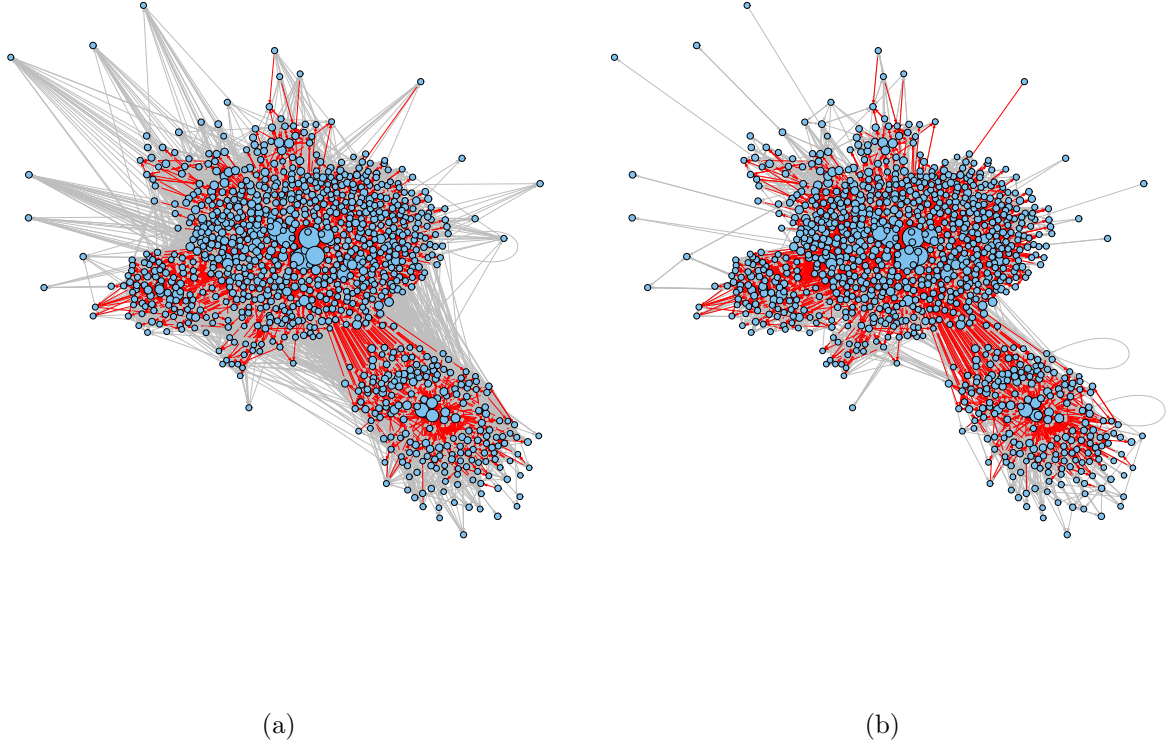


Figure 1: The hyperlink network (a) and the clickstream network (b). The size of nodes denotes the log values of traffic. Edges that are common to both networks are drawn in red. The node layout method is Fruchterman-Reingold, and for ease of comparison, the two networks both have the node layout that was computed for the clickstream network.

of users sharing similar interests will diffuse on the network and label all visited websites until coming to a website that has been “occupied” by another group of users. The detected communities therefore illustrate the preferences of different users. In particular, the simulation works as follows. Each website is initially assigned a unique label and then at every step of the simulation, each website adopts the most popular label in its neighbourhood. The process continues until there is no further change in labels, and websites with the same label are clustered into a community. For example, website 2 in the example network shown in Fig.2 is visited by users from three websites, 0, 1, and 6. After several steps of simulation, website 6 is “occupied” by users coming from website 1, therefore 2 will adopt the same label as 6 and 1. Eventually, websites 1, 6, 2 are clustered into a community, with website 0 belonging to another.

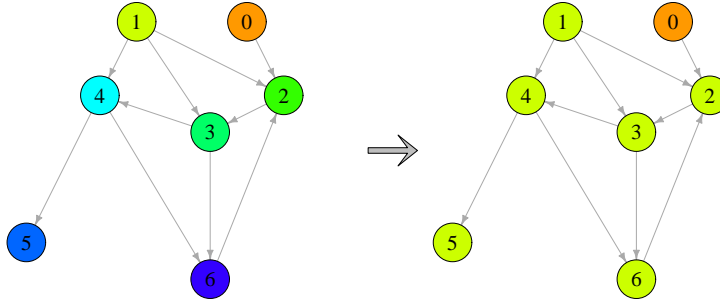


Figure 2: An example showing how the simulation works. In the first step, each website is assigned a unique label. Then, each website adopts the most popular label in its neighbourhood. When the simulation stops, the websites with the same label are clustered into the same community.

Table 1: Descriptive statistics of the clickstream network and its six communities. Clickstream is measured in number of unique visitors)

Community	N of websites	N of edges	Density	Total daily clickstream
Polish Community	4	12	1	2.97×10^6
Korean Community	15	114	0.543	2.29×10^6
Russian Community	28	217	0.287	8.66×10^7
Japanese Community	87	899	0.120	2.33×10^8
Chinese Community	201	2,058	0.052	7.94×10^8
The rest of the world	645	7,695	0.019	4.32×10^9
Total	980	12,008	0.013	5.62×10^9

The simulation identified six communities from the clickstream network (Fig.3), which broadly coincided with visual clusterings provided by the Fruchterman-Reingold algorithm [6]. Websites in each community generally share the same language (with an accuracy rate of 96%, validated by human coding), and we therefore label the communities according to these languages: Polish community, Korean community, Russian community, Japanese community, Chinese community, and The rest of the world. As indicated by Table 1, The rest of the world is the largest, accounting for 645 sites (66% of the total) and 7,695 links (64% of the total) ¹, while the smallest (Polish) consists of only 4 websites.

¹Over 85% of websites within this community are English websites.

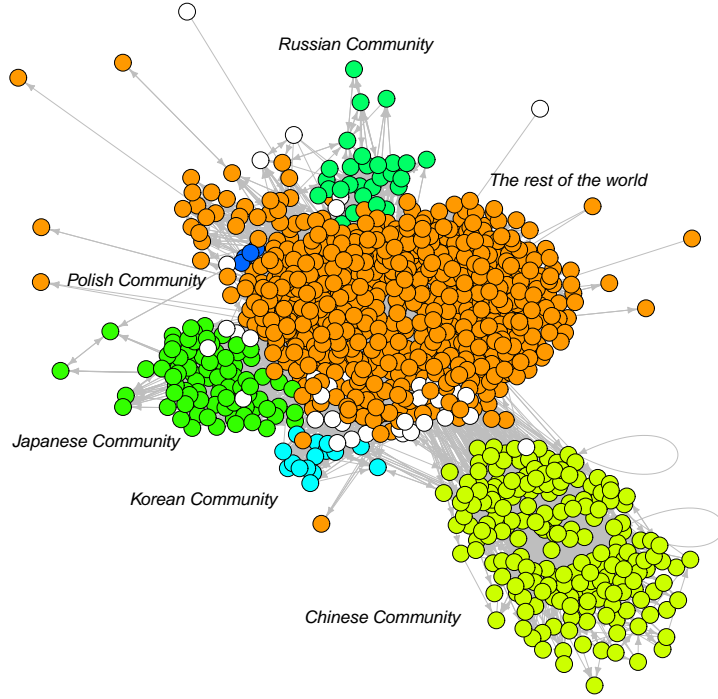


Figure 3: Six language-based communities detected from the clickstream network. The websites in different communities are shown in different colors. The nodes that are assigned to an incorrect community by the label propagation algorithm are plotted in white.

We conduct the simulation on the hyperlink network, but found that all websites were clustered into a single community. The simulation therefore provides compelling evidence that webmasters and users exhibit very different preferences. The hyperlink network is constructed by webmasters, who link their websites to those they think visitors will also be interested in. If users followed the pathways set up by webmasters there would be a very high level of diffusion, as seen in the simulation conducted on the hyperlink network. However, the communities identified in the clickstream network suggest that the linking structure set up by the webmasters does not meet the requirements of users, who in fact prefer to navigate within local, language-based fragments of the Web. We call this behavior “preferential navigation”.

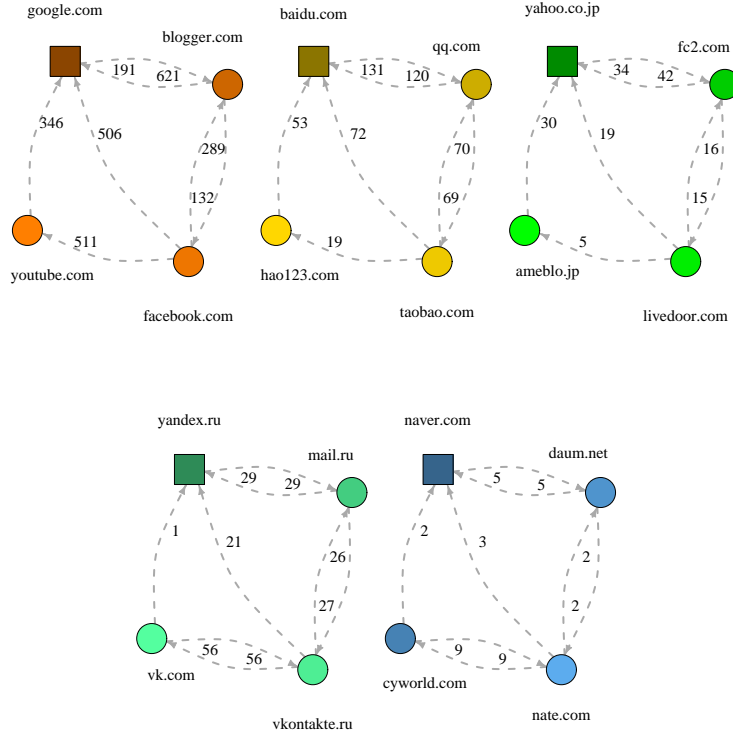


Figure 4: The “popular pathways” in different communities. The squares shows the search engines as starting points and the circles denote other websites on the pathways. The weights on clickstreams indicate traffic measured in millions of unique users. The pathway colors correspond to the colors of communities in Fig.3.

3.3 Understanding web surfing behavior: Preferential navigation driven by local search engines

In the previous section we showed that preferential navigation leads to the creation of language-based website communities. It was mentioned earlier that our clickstream data do not allow us to know the exact process by which a user visits two websites i and j in succession. There are three possibilities: (1) follow a hyperlink from i to j ; (2) enter the URL for j directly into the browser after visiting i (or equivalently, follow a bookmark for j); or (3) navigate from i to a search engine, and then navigate to j by clicking on a search result. We have already shown that (1) does not appear to account for a large proportion of clickstreams. The question we pose in the present section is: how important are search engines in user navigation, and what is their role in enabling preferential navigation?

To answer this question, we examined the role of search engines in the clickstream network.

Firstly, we divided the clickstream network into six sub-networks on the basis of language (this was done manually in light of the fact that the label propagation algorithm had a 4% error rate), and for each sub-network calculated three centrality measures: degree, betweenness, and closeness.² According to [5], these quantities reflect the importance of a node in different aspects: degree indicates the activity of a node, betweenness is a measure of a node’s ability to control the flow in the network, and closeness shows a node’s efficiency in resource transmission. We find that the five search engines, google.com, baidu.com, yahoo.co.jp, yandex.ru, and naver.com, have the highest values in terms of all three centrality measures. This finding clearly points to the predominant role of search engines in facilitating navigation.

The above analysis only reveals the static, structural importance of search engines in the clickstream network; to further investigate the role of search engines in navigation, we compared the clickstreams driven by these five search engines with the clickstreams moderated by hyperlinks. The five search engines moderate over 55% of total clickstreams, with google.com moderating over a half of the clickstreams, followed by baidu.com, yahoo.co.jp, yandex.ru, and naver.com. As mentioned above, hyperlinks only moderate 33% of all the clickstreams. Thus it is reasonable to conclude that users rely more on search engines than hyperlinks in surfing the Web.

At this stage, it is natural to ask exactly how these search engines drive clickstreams? To address this question, we introduce a novel approach for analyzing clickstreams called “popular-pathway-analysis”. This approach is inspired by the “maximizing chain strengths analysis” used in studies of food webs [7]. In each of the five communities, we start from a website ranking 1st in traffic according to Alexa’s statistics. Then in each of the following steps, we choose the strongest (with the largest weight) outbound clickstream. We stop at the fourth step and draw the clickstream sub-network comprising the websites included in the selected clickstreams, which shows the “most likely pathways” a typical user in the communities may take. We found very similar circulations of clickstream in the five communities examined. In these circulations, users start from a “local” search engine³, and return to it repeatedly after

²The Polish community was excluded due to its small size.

³Search engines that index a fragmented part of the Web and moderate clickstreams within a particular community in the clickstream network.

visiting other websites. We contend that this pattern relates to preferential navigation. As local search engines only index a fragmented part of the Web and has a language preference in recommending documents (e.g., baidu.com is more likely to recommend Chinese webpages), successions driven by local search engines are likely to inherit their preferences.

4 Discussion and Conclusion

We present evidence of a mismatch between clickstreams and hyperlinks, and find that in visiting websites, users create their own pathways instead of following hyperlinks passively. We contend that this mismatch originates from the different preferences of webmaster and user: the former set up links to connect to each other's website collaboratively, leading to a highly connected hyperlink network, while the latter use local search engines to guide preferential navigation, which eventually results in a more fragmented, clustered network. It should be noted that although we focus on language in the current study, there are other factors (e.g., culture, economics) that shape clickstream flows.

The findings in this study bring challenges to several areas of web development. For example, hyperlink-based ranking algorithms may provide biased estimates of website relevance [13, 10] since they are derived from the hyperlink network structure created by webmasters, and our research has shown that web users do not tend to follow hyperlinks. Another challenge presented by preferential navigation is: how can search engines become successful in more than one language-based community?

What will happen to the mismatch between clickstreams and hyperlinks as the Web evolves? We believe that as the Web becomes increasingly intelligent, the hyperlink structure will gradually adapt to the clickstream structure, leading to a decrease of the mismatch. In Web 1.0, webmasters are the major constructors of the hyperlinks, and their limited information on user preferences leads to the mismatch. In the era of Web 2.0, users are able to hyperlink by themselves, for example linking from Facebook homepages to blogs. By encouraging users to contribute to the content of websites, webmasters incorporate user preference in setting up hyperlinks, and hence decrease the level of the mismatch [9]. We can imagine that in the era

of Web 3.0, as suggested by Tim Berners-Lee [17], the mismatch between clickstreams and hyperlinks will continue to decrease, as the Web will be able to analyze users' historical surfing records and recommend appropriate websites, leading to the creation of hyperlinks that are more consistent with user web surfing preferences.

Acknowledgements

L. W. thanks Lexing Xie, Paul Thomas, and Hai Liang for providing comments on an earlier version of this paper.

References

- [1] R. Ackland and M. O'Neil. Online collective identity: The case of the environmental movement. *Social Networks*, 2011.
- [2] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of the $1/f$ noise. *Physical Review Letters*, 59(4):381–384, 1987.
- [3] P. Bak, C. Tang, K. Wiesenfeld, et al. Self-organized criticality. *Physical review A*, 38(1):364–374, 1988.
- [4] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.
- [5] L. Freeman. Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1979.
- [6] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21(11):1129–1164, 1991.
- [7] D. Garlaschelli, G. Caldarelli, and L. Pietronero. Universal scaling relations in food webs. *Nature*, 423(6936):165–168, 2003.
- [8] M. Hindman, K. Tsioutsoulouklis, and J. Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *annual meeting of the Midwest Political Science Association*, volume 4, pages 1–33. Citeseer, 2003.

- [9] D. H. Kim, V. Atluri, M. Bieber, N. Adam, and Y. Yesha. A clickstream-based collaborative filtering personalization model: towards a better performance. In *Proceedings of the 6th annual ACM international workshop on Web information and data management*. Association for Computing Machinery, 2004.
- [10] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.
- [11] M. R. Meiss, B. Goncalves, J. J. Ramasco, A. Flammini, and F. Menczer. Agents, bookmarks and clicks: A topical model of web navigation. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, 2010.
- [12] M. R. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Ranking web sites with real user traffic. In *Proceedings of the international conference on Web search and web data mining*, 2008.
- [13] L. Page, S. Brin, R. Motwani, and T. Winograd. Pagerank: Bringing order to the web. Available at: www.pcd.stanford.edu/~page/papers/pagerank. Accessed: January, 29:2001, 1997.
- [14] F. Qiu, Z. Liu, and J. Cho. Analysis of user web traffic with a focus on search activities. In *Proceedings of the International Workshop on the Web and Databases*, 2005.
- [15] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.
- [16] G. Ritzer and N. Jurgenson. Production, consumption, prosumption. *Journal of Consumer Culture*, 10(1):13, 2010.
- [17] N. Shadbolt, W. Hall, and T. Berners-Lee. The semantic web revisited. *Intelligent Systems, IEEE*, 21(3):96–101, 2006.
- [18] D. Tapscott and A. Williams. *Wikinomics: How mass collaboration changes everything*. Portfolio Trade, 2008.
- [19] A. Toffler, W. Longul, and H. Forbes. *The third wave*. Bantam Books New York, 1981.